Ligand-Based Virtual Screening with Co-regularised Support Vector Regression

Conference Paper · December 2016

DOI: 10.1109/ICDMW.2016.0044

5 authors, including:

Michael Kamp
Monash University (Australia)
27 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Martin Vogt
University of Bonn
73 PUBLICATIONS **1,226** CITATIONS
SEE PROFILE

Ligand-Based Virtual Screening with Co-Regularised Support Vector Regression

Katrin Ullrich
University of Bonn
ullrich@iai.uni-bonn.de

Michael Kamp
University of Bonn
Fraunhofer IAIS
michael.kamp@iais.fhg.de

Thomas Gärtner
University of Nottingham
thomas.gaertner@nottingham.ac.uk

Martin Vogt
University of Bonn
B-IT, LIMES Program Unit
martin.vogt@bit.uni-bonn.de

Stefan Wrobel
University of Bonn
Fraunhofer IAIS
stefan.wrobel@iais.fhg.de

***Abstract*—We consider the problem of ligand affinity prediction as a regression task, typically with few labelled examples, many unlabelled instances, and multiple views on the data. In chemoinformatics, the prediction of binding affinities for protein ligands is an important but also challenging task. As protein-ligand bonds trigger biochemical reactions, their characterisation is a crucial step in the process of drug discovery and design. However, the practical determination of ligand affinities is very expensive, whereas unlabelled compounds are available in abundance. Additionally, many different vectorial representations for compounds (molecular fingerprints) exist that cover different sets of features. To this task we propose to apply a co-regularisation approach, which extracts information from unlabelled examples by ensuring that individual models trained on different fingerprints make similar predictions. We extend support vector regression similarly to the existing co-regularised least squares regression (CoRLSR) and obtain a co-regularised support vector regression (CoSVR). We empirically evaluate the performance of CoSVR on various protein-ligand datasets. We show that CoSVR outperforms CoRLSR as well as existing state-of-the-art approaches that do not take unlabelled molecules into account. Additionally, we provide a theoretical bound on the Rademacher complexity for CoSVR.**

I. Introduction

The interactions and bindings of proteins with other molecules are central to most functional processes in biological organisms. Among others, proteins serve as transporters, ion channels, hormones, enzymes, or as transcription factors responsible to regulate processes in the cell. As such modifying the behaviour of a cell by small molecules that bind to specific proteins is a central paradigm for drug-based treatment of diseases. A core task of pharmaceutical research is the drug discovery process, i.e., the identification of novel molecules binding to specific target proteins. As starting point for drug discovery compound databases contain large numbers of molecules that potentially might bind to a protein target of interest. These targets are typically implicated with a disease with the ultimate goal to develop a novel drug treatment. The computational screening of molecular databases for binding partners is called *virtual screening*. Various approaches to virtual screening exist that can be divided into structure-based and ligand-based approaches. Structure-based approaches aim at modelling the *docking* behaviour of molecules on target proteins and require accurate 3-dimensional structural information of the proteins and molecules, as well as knowledge of the binding site. This information, however, is only available for a small set of potential targets. Machine learning methods have been applied successfully in this field [Cherkasov et al., 2014, Ain et al., 2015], but docking models still require expert knowledge of the chemical processes. Ligand-based approaches, on the other hand, try to model the binding affinity to one particular target protein based only on features of the molecules, not taking into account any structural information about the target protein. These methods are oblivious to the actual docking behaviour of molecules and instead learn the binding affinities of novel molecules based on a training set of compounds for which the binding affinity to the respective target protein is known. The co-regularisation approaches presented in this paper are ligand-based virtual screening methods.

The strength of the protein-compound binding interaction is characterised by the so-called real-valued *binding affinity*. A common affinity measure is the *dissociation constant* K_d. If it exceeds a certain limit the small compound is called a *ligand* of the protein. Ligand-based classification models can be trained to distinguish between ligands and non-ligands of the considered protein (e.g., with support vector machines [Geppert et al., 2009]). Since the classification approach represents a severe simplification of the biological reality, we want to predict the strength of binding using regression techniques from machine learning. Although other approaches like *neural networks* have been applied [Myint et al., 2012], *support vector regression* (SVR) is the state-of-the-art method for affinity prediction studies (e.g., [Sugaya, 2014]). In the context of affinity prediction we will use the name ligand for all potential compounds.

For the prediction of affinities one is typically faced with the following practical scenario: for a given protein, only few ligands with experimentally identified affinity values are available. In contrast, the number of synthesisable compounds gathered in molecular databases (such as ZINC, BindingDB,

ChEMBL[1]) that can be used as unlabelled instances for learning is huge. Furthermore, different free or commercial vectorial representations for molecular compounds exist, the so-called molecular fingerprints [Bender et al., 2009]. Originally, each fingerprint type was designed towards a certain learning purpose and, therefore, comprises a characteristic collection of physico-chemical or structural molecular features, for example, predefined key properties (Maccs Keys fingerprint) or listed subgraphs patterns (ECFP fingerprints). Hence, we are confronted with the question of which data representation to choose for the affinity prediction task.

Naturally, it is possible to test and compare different fingerprints [Geppert et al., 2009] or perform preprocessing feature selection and recombination steps on multiple fingerprints [Nisius and Bajorath, 2010] for virtual screening tasks. Furthermore, attempts to utilise multiple fingerprints for one prediction task can be found in the literature [Qiu and Lane, 2008, Ullrich et al., 2016]. However, none of these approaches include unlabelled compounds in the affinity prediction task. Therefore, we propose to apply co-regularisation for regression in order to take profit from both a large number of unlabelled examples and multiple fingerprints without the necessity to choose for one. The semi-supervised *co-regularised least squares regression* (CoRLSR) algorithm of Brefeld et al. [2006] has been shown to outperform single-view *regularised least squares regression* (RLSR) for UCI datasets[2]. Typically, SVR shows very good predictive results having a lower generalisation error compared to RLSR. Moreover, SVR represents the state-of-the-art in affinity prediction (see above). For this reason, we define *co-regularised least squares regression* (CoSVR) as the ε-insensitive loss variant of CoRLSR.

A view on data is a representation of its objects, e.g., with a particular choice of features in $\mathbb{R}^d$. We will see that feature mappings are closely related to the concept of *kernel functions* for which reason the terms *kernel* and *view* are used almost synonymously. Within the research field of *multi-view learning* [Xu et al., 2013], CoSVR and CoRLSR can be assigned to the group of co-training style approaches that simultaneously learn multiple predictors, each related to a view. Co-training style approaches enforce similar outcomes of multiple predictor functions for unlabelled examples, measured with respect to some loss function. In the case of co-regularisation for regression the empirical risks of multiple predictors (*labelled error*) plus an error term for unlabelled examples (*unlabelled error, co-regularisation*) are minimised. In contrast to CoRLSR where the least squares loss is employed we will investigate the ε-insensitive loss function for both labelled and unlabelled error in CoSVR.

The idea for cooperative influence of multiple predictors firstly appeared in a work of Blum and Mitchell [1998] on classification with co-training. Wang et al. [2010] combine the technique of co-training with SVR that is still different from co-regularisation. Analogous to CoSVR, CoRLSR is a semi-supervised and multi-view version of regularised least squares regression that requires the solution of a large system of equations [Brefeld et al., 2006]. A co-regularised version for support vector machine classification SVM-2K already appeared in a work of Farquhar et al. [2006], where the authors define a co-regularisation term via the ε-insensitive loss on labelled examples. It was shown by Sindhwani and Rosenberg [2008] that co-regularised approaches applying the squared loss function for the unlabelled loss can be transformed into a standard SVR optimisation with a particular fusion kernel. A bound on the empirical Rademacher complexity for co-regularised algorithms with Lipschitz continuous loss function for the labelled error and squared loss function for the unlabelled error was proven by Rosenberg and Bartlett [2007].

In the following section, we will present the theoretical foundations of co-regularisation using multiple views. We define CoSVR and variants of its base algorithm in Section III. A Rademacher bound for CoSVR will be proven in Section IV. Subsequently, we provide a practical evaluation of CoSVR for ligand affinity prediction. Finally, Section VI concludes.

II. Preliminaries

A. Kernels and Views

We consider an arbitrary instance space $\mathcal{X}$ and the real numbers as label space $\mathcal{Y}$. We want to learn a function f that predicts a real-valued characteristic of the elements of $\mathcal{X}$. Suppose for training purposes we have sets $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ of labelled and $Z = \{z_1, \ldots, z_m\} \subset \mathcal{X}$ of unlabelled instances at our disposal, where typically $m \gg n$ holds true. With $\{y_1, \ldots, y_n\} \subset \mathcal{Y}$ we denote the respective labels of X. Furthermore, assume the data instances can be represented in M different ways. More formally, for $v \in \{1, \ldots, M\}$ there are functions $\Phi_v : \mathcal{X} \to \mathcal{H}_v$, where $\mathcal{H}_v$ is an appropriate inner product space. Given an instance $x \in \mathcal{X}$, we say that $\Phi_v(x)$ is the *v-th view* of x. If $\mathcal{H}_v$ equals $\mathbb{R}^d$ for some finite dimension d, the intuitive names *(v-th) feature mapping* and *feature space* are used for Φ_v and $\mathcal{H}_v$, respectively. If in the more general case $\mathcal{H}_v$ is a *Hilbert space*, d can even be infinite (see below). For view v the predictor function $f_v : \mathcal{X} \to \mathbb{R}$ is denoted with *(single) view predictor*. View predictors can be learned independently for each view utilising an appropriate regression algorithm like SVR or RLSR. As a special case we consider *concatenated predictors* f_v in Section V where the corresponding view v results from a concatenation of finite dimensional feature representations $\Phi_1, \ldots, \Phi_M$. Having different views on the data, an alternative is to learn M predictors $f_v : \mathcal{X} \to \mathbb{R}$ simultaneously that depend on each other, satisfying an optimisation criterion involving all views at once. Such a criterion could be the minimisation of the labelled error in line with co-regularisation which will be specified in the following subsection. The final predictor f will then be the average of the predictors f_v.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *kernel* if it is symmetric and positive semi-definite. Indeed, for every kernel k there is a feature mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such

[1] zinc.docking.org, www.bindingdb.org, www.ebi.ac.uk/chembl

[2] archive.ics.uci.edu/ml/datasets.html, refer to Newman et al. [1998]

that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) and $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_{\mathcal{H}}$ holds true for all $x_1, x_2 \in \mathcal{X}$ (*Mercer's theorem*). The function k is the reproducing kernel of $\mathcal{H}$, and for $x \in \mathcal{X}$ the mappings $\langle \Phi(x), \Phi(\cdot) \rangle = k(x, \cdot)$ are functions defined on $\mathcal{X}$. Consequently, choosing RKHSs $\mathcal{H}_v$ of multiple kernels k_v as candidate spaces for the predictors f_v the *representer theorem* from Schölkopf et al. [2001] allows for a parameterisation of certain optimisation problems – such as the ones presented below for co-regularisation. A straightforward modification of the representer theorem's proof leads to a representation of the predictors f_v as finite kernel expansion

$$f_v(\cdot) = \sum_{i=1}^{n} \pi_{vi} k_v(x_i, \cdot) + \sum_{j=1}^{m} \pi_{v(j+n)} k_v(z_j, \cdot) \qquad (1)$$

with linear coefficients $\pi_v \in \mathbb{R}^{n+m}$, centered at labelled and unlabelled instances $x_i \in X$ and $z_j \in Z$, respectively.

B. Co-Regularisation for Regression

In order to solve a regression task in the presence of multiple views $v = 1, \ldots, M$, the approach of co-regularisation is to jointly minimise two error terms involving the predictor functions $f_1, \ldots, f_M$. Firstly, every view predictor f_v is intended to have a small training error with respect to labelled examples, i.e., a small labelled error. Secondly, the difference between pairwise view predictions over unlabelled examples should preferably be small, implicating a small unlabelled error. This can be formulated as a *co-regularised empirical risk minimisation* problem

$$\min_{f_v \in \mathcal{H}_v} \quad \sum_{v=1}^{M} \left(\frac{\nu_v}{2} \|f_v\|^2 + \sum_{i=1}^{n} \ell^L(f_v(x_i), y_i) \right) \qquad (2)$$
$$+\lambda \sum_{u,v=1}^{M} \sum_{j=1}^{m} \ell^U(f_u(z_j), f_v(z_j)),$$

where $\nu_v, \lambda \geq 0$ are trade-off parameters and the added norm terms $\|f_v\|$ prevent overfitting. The loss functions ℓ^L and ℓ^U specify the empirical risk and the co-regularisation term. In the case of $\ell^L = \ell^U$ being the squared loss, Problem (2) is known as *co-regularised least squares regression* (CoRLSR). Brefeld et al. [2006] found a closed form solution for CoRLSR as linear system of equations in $M(n+m)$ variables.

C. Notation

At the end of this section we fix some variables and symbols to facilitate the presentation of CoSVR in the following section. The kernel matrices $K_v = \{k_v(x_i, x_j)\}_{i,j=1}^{n+m}$ are the *gram matrices* of the v-th view kernel k_v over labelled and unlabelled examples and have decompositions into L_v and U_v as follows

$$K_v = \begin{pmatrix} L_v \\ U_v \end{pmatrix} \begin{matrix} \in \mathbb{R}^{n \times (n+m)} \\ \in \mathbb{R}^{m \times (n+m)} \end{matrix} .$$

Using L_v and U_v, the predictions of the v-th view predictor on labelled and unlabelled examples can be written as $L_v \pi_v = (f_v(x_1), \ldots, f_v(x_n))^T$ and $U_v \pi_v = (f_v(z_1), \ldots, f_v(z_m))^T$, respectively. With $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and $Y^0 = (y_1, \ldots, y_n, 0, \ldots, 0)^T \in \mathbb{R}^{n+m}$ we denote the label vector and a zero-extension of it. The vector $\mathbb{1}_{(a,b)} \in \mathbb{R}^{n+m}$ is a concatenation of $(a, \ldots, a)^T \in \mathbb{R}^n$ and $(b, \ldots, b)^T \in \mathbb{R}^m$. At some places, we will abbreviate $v \in \{1, \ldots, M\}$ and $(u, v) \in \{1, \ldots, M\}^2$ with $v \in [\![M]\!]$ and $(u, v) \in [\![M]\!]^2$, respectively. Finally, the ε-insensitive loss with parameter ε is defined as $\ell_\varepsilon(y, y') := \max\{0, |y - y'| - \varepsilon\}$, $y, y' \in \mathcal{Y}$.

III. Co-Regularised Support Vector Regression

A. The Base Algorithm

We introduce CoSVR as the co-regularisation Problem (2) with ε-insensitive loss function in the labelled and unlabelled error term.

Definition 1. *For $v \in \{1, \ldots, M\}$ let $\mathcal{H}_v$ be RKHSs. The optimisation problem*

$$\min_{f_v \in \mathcal{H}_v} \sum_{v=1}^{M} \left(\frac{\nu_v}{2} \|f_v\|^2 + \sum_{i=1}^{n} \ell_{\varepsilon^L}(y_i, f_v(x_i)) \right) \qquad (3)$$
$$+\lambda \sum_{u,v=1}^{M} \sum_{j=1}^{m} \ell_{\varepsilon^U}(f_u(z_j), f_v(z_j))$$

is called co-regularised support vector regression (CoSVR), where $\varepsilon^L, \varepsilon^U, \nu_v$, and $\lambda \geq 0$ are algorithm parameters.

Note that the loss function parameters ε^L and ε^U can have different values. In the following, we present a solution for the CoSVR optimisation.

Lemma 1. *Let $\nu_v, \lambda, \varepsilon^L, \varepsilon^U \geq 0$. We use the notation introduced above. In particular, $\pi_v \in \mathbb{R}^{n+m}$ denote the kernel expansion coefficients from Equation (1), whereas $\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n$ and $\gamma_{uv} \in \mathbb{R}^m$ are dual variables. The dual optimisation problem of CoSVR can be formulated as*

$$\max_{\alpha_v, \hat{\alpha}_v, \gamma_{uv}} \sum_{v=1}^{M} \left(-\frac{\nu_v}{2} \pi_v^T K_v \pi_v + \nu_v \pi_v^T Y^0 - \tilde{\pi}_v^T \mathbb{1}_{(\varepsilon^L,\ \varepsilon^U)} \right)$$
$$\textit{where } \pi_v = \frac{1}{\nu_v} \begin{pmatrix} \alpha_v - \hat{\alpha}_v \\ \sum_{u=1}^{M} \gamma_{uv} - \sum_{u=1}^{M} \gamma_{vu} \end{pmatrix}, \ \tilde{\pi}_v = \begin{pmatrix} \alpha_v + \hat{\alpha}_v \\ \sum_{u=1}^{M} \gamma_{uv} \end{pmatrix}$$
$$\textit{s.t. } \begin{Bmatrix} 0_n & \leq & \alpha_v, \hat{\alpha}_v & \leq & 1_n \\ 0_m & \leq & \gamma_{uv} & \leq & \lambda 1_m \end{Bmatrix}_{v \in [\![M]\!],\ (u,v) \in [\![M]\!]^2} .$$

Proof. We showed above that the predictors f_v, $v \in \{1, \ldots, M\}$, have a representation according to Equation (1) as kernel expansion in π_v. With the slack variables $\xi_v, \hat{\xi}_v \in \mathbb{R}^n$,

and $\zeta_{uv} \in \mathbb{R}^m$ we can reformulate the kernelised version of CoSVR as

$$\min_{\pi_v} \sum_{v=1}^{M} \left(\frac{\nu_v}{2} \pi_v^T K_v \pi_v + (\xi_v + \hat{\xi}_v)^T 1_n + \lambda \sum_{u=1}^{M} \zeta_{uv}^T 1_m \right)$$
$$\text{s.t.} \begin{Bmatrix} Y - L_v\pi_v & \leq & \varepsilon^L 1_n + \xi_v \\ L_v\pi_v - Y & \leq & \varepsilon^L 1_n + \hat{\xi}_v \\ U_u\pi_u - U_v\pi_v & \leq & \varepsilon^U 1_m + \zeta_{uv} \\ \xi_v, \hat{\xi}_v & \geq & 0_n \\ \zeta_{uv} & \geq & 0_m \end{Bmatrix}_{v \in [\![M]\!],\ (u,v) \in [\![M]\!]^2},$$

where $\pi_v \in \mathbb{R}^{n+m}$. Introducing Lagrangian multipliers $\alpha_v, \hat{\alpha}_v, \gamma_{uv}, \beta_v, \hat{\beta}_v$, and δ_{uv} for the constraints in the order of appearance in Problem (3), we obtain its Lagrangian $L =$

$$\sum_{v=1}^{M} \Bigg(\frac{\nu_v}{2} \pi_v^T K_v \pi_v + (\xi_v + \hat{\xi}_v)^T 1_n + \lambda \sum_{u=1}^{M} \zeta_{uv}^T 1_m$$
$$+ \alpha_v^T (Y - L_v\pi_v - \varepsilon^L 1_n - \xi_v)$$
$$+ \hat{\alpha}_v^T (L_v\pi_v - Y - \varepsilon^L 1_n - \hat{\xi}_v) - \beta_v^T \xi_v - \hat{\beta}_v^T \hat{\xi}_v$$
$$+ \sum_{u=1}^{M} \gamma_{uv}^T (U_u\pi_u - U_v\pi_v - \varepsilon^U 1_m - \zeta_{uv}) - \sum_{u=1}^{M} \delta_{uv}^T \zeta_{uv} \Bigg).$$

The partial derivatives of L with respect to ξ_v, $\hat{\xi}_v$, and ζ_{uv} set to zero lead us to $L =$

$$\sum_{v=1}^{M} \Big(\frac{\nu_v}{2} \pi_v^T K_v \pi_v + (\alpha_v - \hat{\alpha}_v)^T Y$$
$$- (\alpha_v + \hat{\alpha}_v)^T \varepsilon^L 1_n - \sum_{u=1}^{M} \gamma_{uv}^T\ \varepsilon^U 1_m$$
$$- (\alpha_v - \hat{\alpha}_v)^T L_v\pi_v - \sum_{u=1}^{M} (\gamma_{uv} - \gamma_{vu})^T U_v\pi_v \Bigg).$$

and the box constraints $0_n \leq \alpha_v, \hat{\alpha}_v \leq 1_n$ as well as $0_m \leq \zeta_{uv} \leq \lambda 1_m$. Finally, the stationarity of the partial derivatives $\partial L / \partial \pi_v$ implies the relation

$$\pi_v = \frac{1}{\nu_v} \begin{pmatrix} \alpha_v - \hat{\alpha}_v \\ \sum_{u=1}^{M} \gamma_{uv} - \sum_{u=1}^{M} \gamma_{vu} \end{pmatrix}$$

of primal and dual variables. With the substitution of π_v (and $\tilde{\pi}_v$) into L we finally obtain the desired dual objective. □

The dual optimisation problem in Lemma 1 is a quadratic program (QP) that can be solved with standard QP solvers. We will refer to the CoSVR optimisation in Lemma 1 as the base algorithm.

B. Modifications of the Base Algorithm

The optimisation in Lemma 1 depends on $2Mn + M^2m$ variables, where $m \gg n$. If the number of views M and the number of unlabelled examples m are big, the CoSVR base algorithm might cause problems with respect to running time because of the large number of resulting variables. In order to reduce this number, we define modified versions of base CoSVR. We denote the variant with a modification in the unlabelled error with CoSVR$_{mod}$ and in the labelled error with CoSVR^{mod}.

1) Modification of the Co-Regularisation: The unlabelled error term bounds the pairwise distances of view predictions, whereas now in CoSVR$_{mod}$ only the disagreement between predictions of a view and the average prediction of the residual views will be considered.

Definition 2. *We consider RKHSs $\mathcal{H}_1, \ldots, \mathcal{H}_M$ as well as constants $\varepsilon^L, \varepsilon^U, \nu_v, \lambda \geq 0$. The co-regularised support vector regression problem with modified constraints for the unlabelled examples (CoSVR$_{mod}$) is defined as*

$$\min_{f_v \in \mathcal{H}_v} \sum_{v=1}^{M} \left(\frac{\nu_v}{2} \|f_v\|^2 + \sum_{i=1}^{n} \ell_{\varepsilon^L}(y_i, f_v(x_i)) \right)$$
$$+ \lambda \sum_{v=1}^{M} \sum_{j=1}^{m} \ell_{\varepsilon^U}\left(f_v^{\text{avg}}(z_j),\ f_v(z_j)\right),$$

where

$$f_v^{\text{avg}} := \frac{1}{M-1} \sum_{u=1}^{M, u \neq v} f_u$$

is the average of view predictors besides view v.

Lemma 2. *Let $\nu_v, \lambda, \varepsilon^L, \varepsilon^U \geq 0$. We use dual variables $\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n$ and $\gamma_v, \hat{\gamma}_v \in \mathbb{R}^m$, and define*

$$\gamma_v^{\text{avg}} := \frac{1}{M-1} \sum_{u=1}^{M, u \neq v} \gamma_u \quad \text{and} \quad \hat{\gamma}_v^{\text{avg}} := \frac{1}{M-1} \sum_{u=1}^{M, u \neq v} \hat{\gamma}_u$$

analogous to f_v^{avg}. The CoSVR$_{mod}$ dual optimisation problem can be written as

$$\max_{\alpha_v, \hat{\alpha}_v, \gamma_v, \hat{\gamma}_v} \sum_{v=1}^{M} \left(-\frac{\nu_v}{2} \pi_v^T K_v \pi_v + \nu_v \pi_v^T Y^0 - \tilde{\pi}_v^T \mathbb{1}_{(\varepsilon^L,\ \varepsilon^U)} \right)$$
$$\text{where } \pi_v = \frac{1}{\nu_v} \begin{pmatrix} \alpha_v - \hat{\alpha}_v \\ (\gamma_v - \gamma_v^{\text{avg}}) - (\hat{\gamma}_v - \hat{\gamma}_v^{\text{avg}}) \end{pmatrix},$$
$$\tilde{\pi}_v = \begin{pmatrix} \alpha_v + \hat{\alpha}_v \\ \gamma_v + \hat{\gamma}_v \end{pmatrix}$$
$$\text{s.t.} \begin{Bmatrix} 0_n & \leq & \alpha_v, \hat{\alpha}_v & \leq & 1_n \\ 0_m & \leq & \gamma_v, \hat{\gamma}_v & \leq & \lambda 1_m \end{Bmatrix}_{v \in [\![M]\!]}.$$

Proof. The proof is analogous to the one of Lemma 1 with slack variables $\xi_v, \hat{\xi}_v, \zeta_v, \hat{\zeta}_v$, as well as dual variables $\alpha_v, \hat{\alpha}_v, \beta_v, \hat{\beta}_v, \gamma_v, \hat{\gamma}_v, \delta_v$, and finally, $\hat{\delta}_v$. □

In the base CoSVR versions the semi-supervision is realised with proximity constraints on pairs of view predictions. We show in the following lemma that the constraints of the closeness of one view prediction to the average of the residual predictions implies a closeness of every pair of predictions.

Lemma 3. *For appropriate choices of ε^U the unlabelled error of CoSVR$_{mod}$ is an upper bound of the unlabelled error of base CoSVR.*

Proof. We consider the settings of Lemma 1 and Lemma 2. In the case of $M = 2$

$$\sum_{v=1}^{M}\sum_{j=1}^{m}\max\{0, |f_v^{\text{avg}}(z_j) - f_v(z_j)| - \varepsilon^U\}$$
$$= \sum_{u,v=1}^{M}\sum_{j=1}^{m}\max\{0, |f_u(z_j) - f_v(z_j)| - \varepsilon^U\}$$

holds true which shows the claim. Now let $M > 2$. Because of the definition of the ε-insensitive loss and the role of ζ in the proof of Lemma 2 we know that

$$|f_v(z_j) - f_v^{\text{avg}}(z_j)| \leq \varepsilon^U + \max_{v\in\{1,\dots,M\}}\{|\zeta_v|, |\hat{\zeta}_v|\}$$

for all $v \in \{1,\dots,M\}$. We denote $c_j := \max_{v\in\{1,\dots,M\}}\{|\zeta_v|, |\hat{\zeta}_v|\}$ and, hence, $|f_v(z_j) - f_v^{\text{avg}}(z_j)| \leq \varepsilon^U + c_j$. Now we conclude for $j \in \{1,\dots,m\}$ and all $(u,v) \in \{1,\dots,M\}^2$ that

$$\begin{aligned}
&|f_u(z_j) - f_v(z_j)| \\
&\leq |f_u(z_j) - f_u^{\text{avg}}(z_j)| + |f_u^{\text{avg}}(z_j) - f_v^{\text{avg}}(z_j)| \\
&\qquad + |f_v^{\text{avg}}(z_j) - f_v(z_j)| \\
&\leq \varepsilon^U + c_j + \tfrac{1}{M-1}|f_v(z_j) - f_u(z_j)| + \varepsilon^U + c_j,
\end{aligned}$$

and therefore,

$$|f_u(z_j) - f_v(z_j)| \leq \frac{2(M-1)}{M-2}(\varepsilon^U + c_j).$$

As a consequence we deduce from

$$\sum_{j=1}^{m}\sum_{v=1}^{M}\ell_{\varepsilon^U}(f_v^{\text{avg}}(z_j), f_v(z_j)) \leq M\sum_{j=1}^{m}c_j = B$$

that

$$\sum_{j=1}^{m}\sum_{u,v=1}^{M}\ell_{\tilde{\varepsilon}}(f_u(z_j), f_v(z_j)) \leq \tfrac{2(M-1)}{M(M-2)}B$$

for $\tilde{\varepsilon} = \frac{2(M-1)}{M-2}\varepsilon^U$, which finishes the proof. □

2) Modification of the Empirical Risk: We can also reduce the number of variables using modified constraints for the empirical risk term.

Definition 3. *The co-regularised support vector regression problem with modified constraints for the labelled examples (CoSVR^{mod}) is defined as*

$$\min_{f_v\in\mathcal{H}_v}\sum_{v=1}^{M}\frac{\nu_v}{2}\|f_v\|^2 + \sum_{i=1}^{n}\ell_{\varepsilon^L}(y_i, f^{\text{avg}}(x_i))$$
$$+\lambda\sum_{u,v=1}^{M}\sum_{j=1}^{m}\ell_{\varepsilon^U}\left(f_u(z_j), f_v(z_j)\right),$$

where now

$$f^{\text{avg}} := \frac{1}{M}\sum_{v=1}^{M}f_v.$$

If we combine the modifications in the labelled and unlabelled error term we obtain a third variant CoSVR$_{mod}^{mod}$.

TABLE I
NUMBER OF VARIABLES, CONSTRAINTS, AND POTENTIAL NON-ZERO VARIABLES FOR DIFFERENT COSVR VERSIONS AND CORLSR. THE RESPECTIVE COSVR^{mod}-METHODS ARE INCLUDED BY CANCELLING THE $\{M\}$-FACTOR.

	variables	**constraints**	**sparsity**
CoSVR	$2\{M\}n$ $+M^2m$	$4\{M\}n$ $+2M^2m$	$\{M\}n$ $+\frac{1}{2}(M^2-M)m$
CoSVR$_{mod}$	$2\{M\}n$ $+2Mm$	$4\{M\}n$ $+4Mm$	$\{M\}n$ $+Mm$
CoRLSR	$Mn + Mm$	0	$Mn + Mm$

Lemma 4. *The dual problems of CoSVR$_{mod}^{mod}$ and CoSVR^{mod} are the ones of the respective duals of CoSVR$_{mod}$ and CoSVR up to a substitution of α_v and $\hat{\alpha}_v$ by view-independent variable vectors $\frac{1}{M}\alpha$ and $\frac{1}{M}\hat{\alpha} \in \mathbb{R}^n$ for all $v \in 1,\dots,M$.*

Proof. The reduction of variables directly follows from the view-independence of the labelled error. □

C. Complexity

The CoSVR variants and CoRLSR mainly differ in the number of applied loss functions and the strictness of constraints. This results in different numbers of variables and constraints in total, as well as potentially non-zero variables (referred to as *sparsity*, compare Table I). All presented problems are convex QPs with positive semi-definite matrices in the quadratic terms. As the number m of unlabelled instances in real-world problems is much bigger than n, the running time of a QP-solver is dominated by the respective second summand in the constraints column of Table I. Because of the ε-insensitive loss the number of actual non-zero variables in the learned model will be even higher for the CoSVR$_{(mod)}^{(mod)}$ variants than the numbers reported in the sparsity column of Table I. In particular, for the modified variants this will allow for a more efficient model storage compared to CoRLSR. Indeed, according to the *Karush-Kuhn-Tucker conditions* (compare for example Boyd and Vandenberghe [2004]), only for active inequality constraints the corresponding dual γ-variables can be non-zero. In this sense the respective unlabelled $z_j \in Z$ are *unlabelled support vectors*. This consideration is also valid for the α-variables and support vectors $x_i \in X$ as we use the ε-insensitive loss for the labelled and unlabelled error in all CoSVR versions.

In the two-view case with $M = 2$ the modified version with respect to the unlabelled error term and the base version fall together. It is still useful, however, to implement it as the modified way. If we additionally exploit that for the dual variables $\gamma_1 = \hat{\gamma}_2$ and $\gamma_2 = \hat{\gamma}_1$ holds true we can reduce the number of variables even further and all numbers in Table I in the second row can be multiplied with $\frac{1}{2}$.

IV. A RADEMACHER BOUND FOR COSVR

By now, we restrict to a two-view setting, i.e., $M = 2$. Similarly to the result of Rosenberg and Bartlett [2007]

we want to prove a bound on the *empirical Rademacher complexity* of CoSVR. Let us define the sum space $\mathcal{H}_\Sigma$

$$\mathcal{H}_\Sigma := \{f : f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}. \quad (4)$$

The empirical Rademacher complexity $\hat{\mathcal{R}}_n$ is a data-dependent measure for the capacity of a function class $\mathcal{H}$ to fit random data [Shawe-Taylor and Christianini, 2004] and is defined as

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}^\sigma \left[\sup_{f\in\mathcal{H}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| : \{x_1, \ldots, x_n\} = X \right].$$

The random data are represented via the Rademacher random variables $\sigma = (\sigma_1, \ldots, \sigma_n)^T$. Additionally, we define a bounded version $\mathcal{H}_\Sigma^\varepsilon$ of the sum space in Equation (4). Therefore, let π_v, K_v, L_v, and U_v be the kernel expansion parameters from Equation (1), the gram matrices over labelled and unlabelled examples, and the upper as well as the lower part of the gram matrices K_v of f_v, $v \in \{1, 2\}$, as defined above. Obviously, a pair (π_1, π_2) represents an element of $\mathcal{H}_\Sigma$. From Lemma 1 and Lemma 2 we see that π_1 and π_2 must be bounded by some constant μ that depends on ν_1, ν_2, λ, and M. Hence, we define

$$\mathcal{H}_\Sigma^\varepsilon := \{(\pi_1, \pi_2) \in \mathcal{H}_\Sigma : -\mu 1_{n+m} \leq \pi_1, \pi_2 \leq \mu 1_{n+m}\}.$$

We consider CoSVR or its variant CoSVR$^{(mod)}$ and prove a bound for the empirical Rademacher complexity of $\mathcal{H}_\Sigma^\varepsilon$. We point out that for two views the modified and base versions with respect to the co-regularisation part fall together. With sparsity again we denote the maximum number of vector entries different from zero.

Lemma 5. *Let $\mathcal{H}_\Sigma^\varepsilon$ be the function space defined above and, w.l.o.g., let $\mathcal{Y} = [-1, 1]$. The empirical Rademacher complexity of CoSVR$^{(mod)}$ can be bounded via*

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^\varepsilon) \leq \frac{2s}{n}\mu(\|L_1\|_\infty + \|L_2\|_\infty), \quad (5)$$

where μ is a constant dependent on the regularisation parameters and s is the sparsity of the kernel expansion vector $(\pi_1, \pi_2) \in \mathcal{H}_\Sigma^\varepsilon$.

Our proof applies Theorem 2 in Rosenberg and Bartlett [2007] and generalises the upper bound of Theorem 3 in Rosenberg and Bartlett [2007].

Proof. At first, we investigate the general usefulness of the empirical Rademacher complexity $\hat{\mathcal{R}}_n$ of $\mathcal{H}_\Sigma^\varepsilon$ in the CoSVR scenario. Therefore, we notice that the ε-insensitive loss function in the labelled error

$$\ell_{\varepsilon^L}(y, f(x)) \quad = \quad \max\{0, |y - (f_1(x) + f_2(x))/2| - \varepsilon^L\}$$

maps into $[0, 1]$ because of the boundedness of $\mathcal{Y}$. Furthermore, it is easy to show that ℓ_{ε^L} is *Lipschitz continuous*, i.e., $|\ell_{\varepsilon^L}(y, x) - \ell_{\varepsilon^L}(y, x')|/|x - x'| \leq C$, for some constant $C > 0$. With similar arguments one can show that the standard ε-insensitive loss function is Lipschitz continuous as well. According to Theorem 2 in Rosenberg and Bartlett [2007], the expected loss $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ \ell_{\varepsilon^L}(f(X), Y)$ can then be bounded by means of the empirical risk and the empirical Rademacher complexity

$$\mathbb{E}_\mathcal{D}\ \ell_{\varepsilon^L}(f(X), Y) \leq \frac{1}{n}\sum_{i=1}^{n} \ell_{\varepsilon^L}(f(x_i), y_i) + 2C\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^\varepsilon) + \frac{2 + 3\sqrt{\ln(2/\delta)/2}}{\sqrt{n}}$$

for every $f \in \mathcal{H}_\Sigma^\varepsilon$ with probability at least $1 - \delta$. We can now reformulate the empirical Rademacher complexity

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^\varepsilon) &= \mathbb{E}^\sigma \left[\sup_{f\in\mathcal{H}_\Sigma^\varepsilon} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| : x_i \in X \right] \\ &= \frac{2}{n}\mathbb{E}^\sigma \left[\sup_{(\pi_1,\pi_2)^T\in\mathcal{K}} \left|\sigma^T(L_1\pi_1 + L_2\pi_2)\right| \right], \end{aligned}$$

where

$$\mathcal{K} := \{(\pi_1, \pi_2)^T \in \mathbb{R}^{2(n+m)} : |\pi_1|, |\pi_2| \leq 1_{n+m}\mu\}.$$

In the expression above, L_1 and L_2 denote the upper parts of the kernel matrices K_1 and K_2, respectively. The kernel expansion $\pi = (\pi_1, \pi_2)$ for the CoSVR$^{(mod)}$ optimisation is bounded because of the box constraints in the respective dual problems. Therefore, π lies in the ℓ_1-ball of dimension s scaled with $s\mu$, i.e., $\pi \in s\mu \cdot B_1$. The dimension s is the sparsity of π, and thus, the number of expansion variables π_{vj} different from zero. From the dual optimisation problem, we know that $s \ll 2(n + m)$. It is a fact (Theorems II.2.3 and II.2.4 in Werner [1995]) that $\sup_{\pi\in s\mu\cdot B_1} |\langle w, \pi\rangle| = s\mu\|w\|_\infty$. Let $L \in \mathbb{R}^{n\times 2(n+m)}$ be the concatenated matrix $L = (L_1|L_2)$, where L_1 and L_2 are the upper parts of the gram matrices K_1 and K_2. From the definition we see that $w = \sigma^T L$ and, hence,

$$\begin{aligned} s\mu\|w\|_\infty &= s\mu\|\sigma^T L\|_\infty \leq s\mu\|\sigma\|_\infty\|L\|_\infty \leq s\mu\|L\|_\infty \\ &= s\mu \max_{i\in[n]} \sum_{j\in[n+m]} \sum_{v=1,2} |k_v(x_i, x_j)|, \end{aligned}$$

where $\|L\|_\infty$ is the *row sum norm* of L. Finally, we obtain for the empirical Rademacher complexity

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^\varepsilon) \leq \frac{2}{n}\mathbb{E}^\sigma s\mu\|L\|_\infty \leq \frac{2s}{n}\mu(\|L_1\|_\infty + \|L_2\|_\infty)$$

which was the desired result. □

V. Empirical Evaluation

In this section we evaluate the performance of CoSVR for predicting the activity values of molecules against a target protein.

Our experiments are performed on 24 datasets, consisting of ligands and their affinity to one particular human protein per dataset, gathered from BindingDB[3]. Every ligand is a single molecule in the sense of a connected graph and is labelled with its affinity value towards the protein target. The ligands are available in the standard molecular fingerprint formats ECFP4, GpiDAPH3, and Maccs Keys.

[3]Binding database, https://www.bindingdb.org/bind/index.jsp

TABLE II
DATASETS WITH NAME, NUMBER OF LIGANDS, AND LABEL RANGE.

Target	Ligands	Range	Target	Ligands	Range
P14091	21	$6.1 - 10.0$	P42574	133	$4.9 - 11.9$
P08311	23	$3.9 - 9.8$	P00740	171	$3.9 - 8.7$
Q16651	23	$4.8 - 7.9$	P07384	189	$3.1 - 10.7$
P07288	28	$7.2 - 9.7$	P07339	197	$4.1 - 11.0$
P04070	31	$3.7 - 7.1$	P08709	249	$3.9 - 9.5$
O60235	41	$5.8 - 7.9$	P43235	252	$3.9 - 11.5$
P03952	76	$3.0 - 9.3$	P00750	268	$2.2 - 9.5$
P23946	90	$5.4 - 8.9$	P07858	278	$3.0 - 10.5$
Q99895	91	$2.7 - 8.0$	P29466	310	$3.1 - 9.8$
P09871	92	$4.8 - 9.0$	P07711	357	$3.9 - 10.6$
P25774	104	$4.3 - 9.8$	P00747	474	$1.9 - 11.0$
P17655	128	$4.8 - 10.8$	P00749	600	$0.3 - 11.1$

We compare CoSVR against co-regularised least squares regression (CoRLSR), as well as a support vector regression on a single view (SVR) in terms of root mean squared error (RMSE). Another natural baseline is to apply an SVR to a new view that is created by concatenating the features of all views (SVR(concat)). We also compare CoSVR against an oracle that chooses the best SVR for each view and each dataset (SVR(best)) by taking the result with the best performance in hindsight.

We consider virtual screening as semi-supervised learning with many unlabelled data instances. Thus, we split each labelled dataset into a labelled (30% of the examples) and an unlabelled part (the remaining 70%). The labelled part is used as training set, whereas the unlabelled part is the test set. Since the co-regularised approaches are semi-supervised, they employ both labelled and unlabelled examples, i.e., they have access to the entire set of unlabelled examples in each fold. The RMSE is measured using 5-fold cross-validation. The parameters for each approach on each dataset are optimised using grid search with 5-fold cross-validation on a sample of the training set—because of the small size of the datasets, using an independent sample for parameter optimisation is practically infeasible.

In Fig. 1 we present the results of CoSVR compared to CoRLSR and various single view approaches for all datasets using the fingerprints GpiDAPH3 and ECFP4. Fig. (a) indicates that CoSVR outperforms CoRLSR on the majority of datasets. It also outperforms SVR on the GpiDAPH3 view (Fig. (b)), the SVR(concat) (Fig. (d)) and the SVR on the ECFP4 view (Fig (c)), the latter only slightly. Fig. (e) indicates that SVR(best) performs better than the other baselines but is still outperformed by CoSVR.

The indications in Fig. 1 are substantiated by a Wilcoxon Signed-Rank test on the results, presented in Table III. In this table, we report the median RMSEs and the test statistics (T and p-value). Results in which CoSVR statistically significantly outperforms the baselines (for a significance level $p < 0.05$) are marked in bold. The test confirms that CoSVR performs statistically significantly better than CoRLSR, as well as an SVR trained on each individual view, the concatenation of views and taking the best single view SVR in hindsight.

Note that the reported results have been obtained with the modified version CoSVR$_{mod}$ of CoSVR. A Wilcoxon Signed-Rank test on the RMSEs of CoSVR in its original formulation and the modified variant CoSVR$_{mod}$ did not find a statistically significant difference in the RMSEs of the methods (with $Z = 87.5$ and $p < 0.5135$).

TABLE III
COMPARING RMSES USING WILCOXON SIGNED-RANK TEST

baseline	Z	p-value
CoRLSR	14.0	$< \mathbf{0.00011}$
SVR(GpiDAPH3)	0.0	$< \mathbf{0.00001}$
SVR(ECFP4)	33.0	$< \mathbf{0.00083}$
SVR(concat)	2.0	$< \mathbf{0.00003}$
SVR(best)	42.5	$< \mathbf{0.00213}$

In Table IV we report the average RMSEs of CoSVR, CoRLSR and the single view baselines for all combinations of the views Maccs Keys, GpiDAPH3, and ECFP4. In terms of average RMSE, CoSVR outperforms the other approaches for the view combination Maccs Keys and GpiDAPH3, as well as GpiDAPH3 and ECFP4. For the views Maccs Keys and ECFP4, CoSVR has lower average RMSE than CoRLSR and the single view SVRs. However, for this view combination, the SVR(best) as well as the SVR(concat) baselines outperform CoSVR.

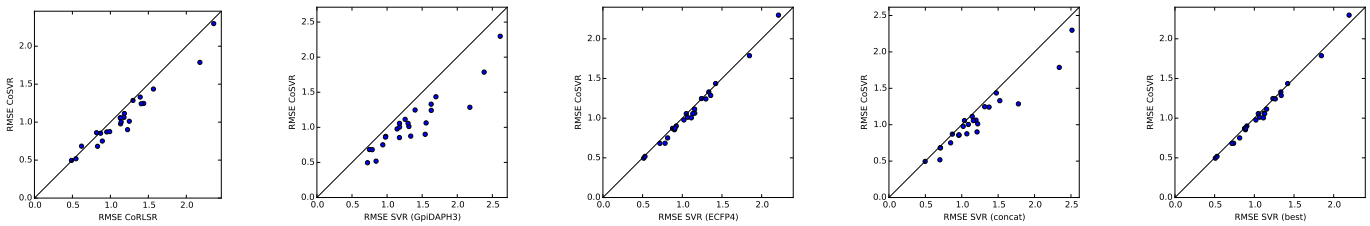
In conclusion, co-regularisation techniques perform better than the state of the art of single view approaches, as well as a concatenation of features from multiple views. In particular, CoSVR outperforms CoRLSR and SVR on all view combinations, as well as SVR(concat) on 2 out of 3 view combinations. Moreover, CoSVR performs as good as the SVR(best) on 2 out of 3 view combinations.

TABLE IV
AVERAGE RMSES FOR ALL COMBINATIONS OF THE FINGERPRINTS MACCS KEYS, GPIDAPH3, AND ECFP4

Method	View Combinations			
	Maccs, ECFP4	Maccs, GpiDAPH3	GpiDAPH3, ECFP4	Maccs, GpiDAPH3, ECFP4
CoSVR	0.994	**1.006**	**1.058**	0.936
CoRLSR	1.015	1.084	1.168	1.084
SVR(view1)	1.007	1.031	1.354	0.953
SVR(view2)	1.011	1.373	1.095	1.501
SVR(view3)	-	-	-	1.114
SVR(concat)	0.938	1.143	1.196	1.157
SVR(best)	**0.929**	1.009	1.085	**0.891**

VI. CONCLUSION

We successfully applied co-regularised support vector regression to the problem of ligand affinity prediction, however, at the cost of solving a more complex optimisation problem, resulting in a higher runtime than single-view approaches. Experiments show that CoSVR outperforms the state-of-the-art approaches in ligand-based virtual screening.



(a) (b) (c) (d) (e)

Fig. 1. Comparison of CoSVR with CoRLSR and single view SVR on 24 datasets using the features GpiDAPH3 and ECFP4 in terms of RMSEs. Each point represents the RMSEs of the two methods compared on one dataset.

REFERENCES

Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Comput. Mol. Sci.*, 2015.

A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.*, 2009.

A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of COLT 1998*, 1998.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient Co-Regularised Least Squares Regression. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. Baskin, M. Cronin, and et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.*, 2014.

J. D. R. Farquhar, H. Meng, S. Szedmak, D. Hardoon, and J. Shawe-Taylor. Two view learning: SVM-2K, Theory and Practice. In *Proceedings of NIPS 2006*, 2006.

H. Geppert, J. Humrich, D. Stumpfe, T. Gärtner, and J. Bajorath. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.*, 2009.

K.-Z. Myint, L. Wang, Q. Tong, and X.-Q. Xie. Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. *Mol. Pharmaceutics*, 2012.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

B. Nisius and J. Bajorath. Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of Hits. *Chem. Biol. Drug Des.*, 2010.

S. Qiu and T. Lane. Multiple Kernel Support Vector Regression for siRNA Efficacy Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

D. S. Rosenberg and P. L. Bartlett. The Rademacher Complexity of Co-Regularized Kernel Classes. In *Proceedings of AISTATS 2007*, 2007.

B. Schölkopf, R. Herbrich, A. J. Smola, and R. Williamson. A Generalized Representer Theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2001.

J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

V. Sindhwani and D. S. Rosenberg. An RKHS for Multi-View Learning and Manifold Co-Regularization. In *Proceedings of ICML 2008*, 2008.

N. Sugaya. Ligand Efficiency-Based Support Vector Regression Models for Predicting Bioactivities of Ligands to Drug Target Proteins. *J. Chem. Inf. Model.*, 2014.

K. Ullrich, J. Mack, and P. Welke. Ligand Affinity Prediction with Multi-Pattern Kernels. *To be published in Proceedings of DS 2016*, 2016.

X. Wang, L. Ma, and X. Wang. Apply semi-supervised support vector regression for remote sensing water quality retrieving. *Proceedings of IGARSS 2010 IEEE International*, 2010.

D. Werner. *Funktionalanalysis*. Springer, 1995.

C. Xu, D. Tao, and C. Xu. A Survey on Multi-view Learning. *Proceedings of CoRR 2013*, 2013.